Name: Satyam Rai
UID: 2019130051
TE COMPS
Batch C

EXPERIMENT 5

(K-Means Clustering)

Aim: To train a machine learning model using K-means clustering algorithm to find the optimal value of number of clusters for the income and house price of people found in the dataset.

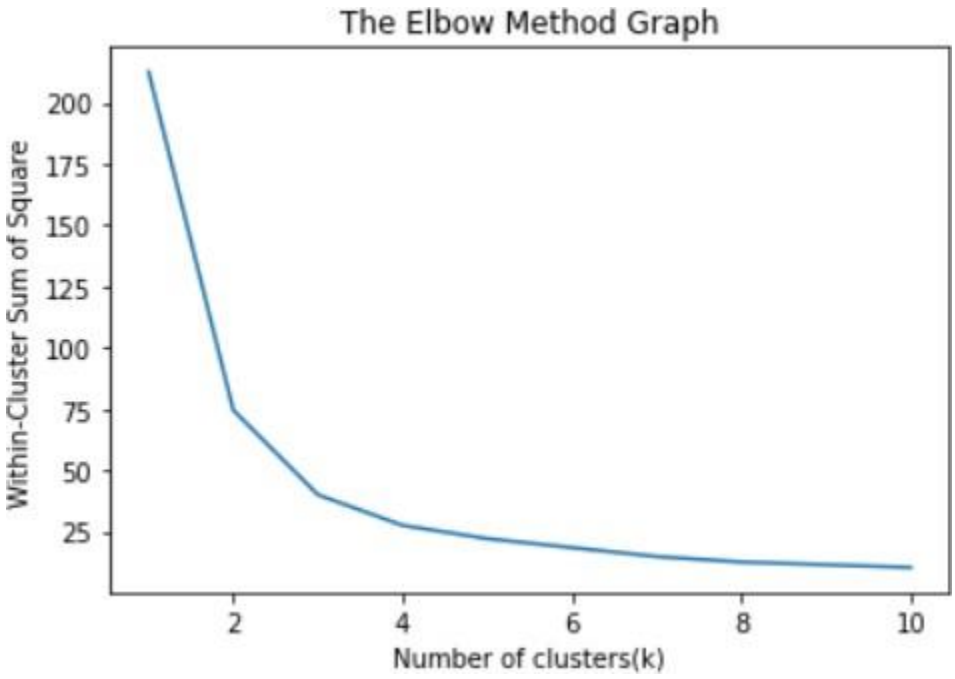
Code:

```
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
from sklearn.cluster import KMeans

df = pd.read_csv('housing.csv')
df = df.iloc[1400:1500]
data_for_clustering = df[["MedInc","MedHouseVal"]]

x = data_for_clustering.values

plt.scatter(data_for_clustering.MedInc.to_list() ,
data_for_clustering.MedHouseVal.to_list())
plt.title("House Prices")
plt.xlabel("Income")
plt.ylabel("House prices")
plt.show()

def get_wcss(X):
    wcss_list= []
    for i in range(1, 11):
        kmeans = KMeans(n_clusters=i, init='k-means++', random_state= 42)
        kmeans.fit(X)
        wcss_list.append(kmeans.inertia_)

    return wcss_list

wcss = get_wcss(x)
```

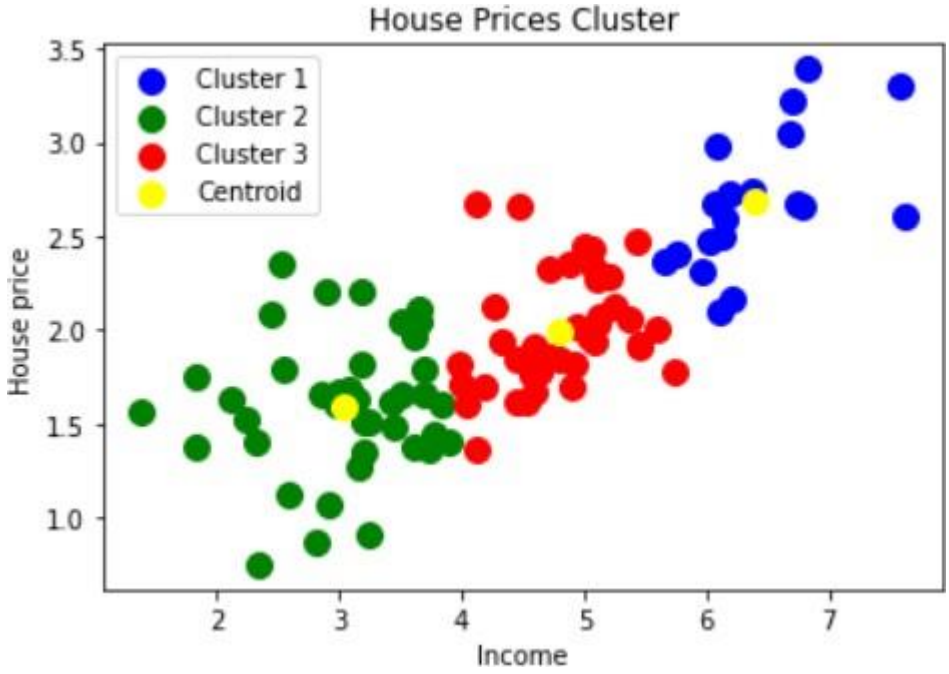
```
print(wcss)
plt.plot(range(1, 11), wcss)
plt.title('The Elbow Method Graph')
plt.xlabel('Number of clusters(k)')
plt.ylabel('Within-Cluster Sum of Square')
plt.show()

def clustering_kmeans(X,k):
    kmeans = KMeans(n_clusters=k, init='k-means++', random_state= 42)
    y= kmeans.fit_predict(X)
    return kmeans,y

k_means, y = clustering_kmeans(x, 3)

plt.scatter(x[y == 0, 0], x[y == 0, 1], s = 100, c = 'blue', label =
'Cluster 1')
plt.scatter(x[y == 1, 0], x[y == 1, 1], s = 100, c = 'green', label =
'Cluster 2')
plt.scatter(x[y== 2, 0], x[y == 2, 1], s = 100, c = 'red', label =
'Cluster 3')
# plt.scatter(x[y == 3, 0], x[y == 3, 1], s = 100, c = 'cyan', label =
'Cluster 4')
plt.scatter(k_means.cluster_centers_[:, 0], k_means.cluster_centers_[:,
1], s = 100, c = 'yellow', label = 'Centroid')
plt.title('House Prices Cluster')
plt.xlabel('Income')
plt.ylabel('House price')
plt.legend()
plt.show()
```

Observation:

WCSS scores for cluster values from 1-10-
[212.67648267190006, 74.57729755995962, 40.00232767936331,
27.520531769942124, 22.102294138198378, 18.442280411841995,
14.809002254054407, 12.558447298241122, 11.467933900351197,
10.353005105846478]



From the above Elbow Method graph which is wcss values plotted against each cluster value k we observe that the optimal value of k is 3.



Conclusion:

In this experiment, the aim was to implement the K-Means clustering algorithm and to run the algorithm on a random dataset which in this case is a dataset containing the income and house price of people. Then using k-means algorithm, for k values 1-10, I calculated the Within-Cluster Sum of Squares(WCSS) values and plotted the Elbow Method graph which helps to find out the optimal number of clusters which comes out to be 3 using the graph. Then using this k value the graph with 3 clusters for income and house price is plotted and all the centroids along with their data points are plotted.